Supplement to the Annals of Applied Statistics
arXiv: math.PR/0000000

SUPPLEMENT TO MODELING SOCIAL NETWORKS FROM SAMPLED DATA*

BY MARK S. HANDCOCK[§] AND KRISTA J. GILE[‡]

University of California - Los Angeles and Nuffield College

Network models are widely used to represent relational information among interacting units and the structural implications of these relations.

Handcock and Gile (2010) develop the conceptual and computational theory for inference based on sampled network information. They first review forms of network sampling designs used in practice. They consider inference from the likelihood framework, and develop a typology of network data that reflects their treatment within this frame. They then develop inference for social network models based on information from adaptive network designs.

They motivate and illustrate these ideas by analyzing the effect of link-tracing sampling designs on a collaboration network.

In this supplement we provide the code used to perform this study. It is written in the `R` statistical language (R Development Core Team, 2007) and is based on `statnet`, an open-source software suite for network modeling (Handcock *et al.*, 2003). We provide documentation with links to the `statnet` website.

The URL for this supplement is `http://lib.stat.cmu.edu/aoas/221/supplement.zip`

1. Likelihood Computations for Exponential Family Models for Networks. Using the notation in Handcock and Gile (2010), consider a parametric model for the random behavior of a network Y depending on a parameter $p-$vector η:

$$P_\eta(Y = y) \qquad \eta \in \Xi \tag{1.1}$$

*This work was supported by NIH awards R01 DA012831 and R01 HD041877, NSF award MMS-0851555, and ONR award N00014-08-1-1015. The authors would like to thank the members of the UW Network Modeling Group (Martina Morris, P.I.), Stephen Fienberg, and the reviewers for their helpful input.

†Submitted December 7, 2007; Revised September 29, 2008

‡Current address: Nuffield College, University of Oxford, New Road, Oxford 0X1 1NF, United Kingdom.

§Current address: Department of Statistics, University of California, Los Angeles CA 90095-1554, USA.

AMS 2000 subject classifications: Primary 91D30, 62D05; secondary 60K35

Keywords and phrases: exponential family random graph model, p^* model, Markov chain Monte Carlo, design-based inference

In the model-based framework, if Y is completely observed inference for η can be based on the likelihood:

$$L[\eta|Y = y] \propto P_\eta(Y = y).$$

While the paper encompasses general models for complete networks, it focuses on exponential family random graph models (ERGM) in the applications. These model the network Y through a p-vector $g(Y)$ of statistics. The canonical exponential family model is

$$P_\eta(Y = y) = \exp\{\eta \cdot g(y) - \kappa(\eta)\} \qquad y \in \mathcal{Y} \tag{1.2}$$

where $\exp\{\kappa(\eta)\} = \sum_{u \in \mathcal{Y}} \exp\{\eta \cdot g(u)\}$ is the familiar normalizing constant associated with an exponential family of distributions (Barndorff-Nielsen, 1978; Lehmann, 1983).

The range of network statistics that might be included in the $g(y)$ vector is vast. We allow the vector $g(y)$ to include covariate information about nodes or edges in the graph in addition to information derived directly from the matrix y itself.

The normalizing constant is usually difficult to compute directly for $\mathcal{Y}$ containing large numbers of networks. Inference for this class of models was considered in the seminal paper by Geyer and Thompson (1992), building on the methods of Frank and Strauss (1986) and other papers cited in Handcock and Gile (2010).

Geyer and Thompson (1992) proposed a stochastic algorithm to approximate maximum likelihood estimates for model (1.2), among other models; this Markov chain Monte Carlo (MCMC) approach forms the basis of the method used in Handcock and Gile (2010).

These methods are implemented in the **statnet** open-source statistical network analysis system (Handcock *et al.*, 2003) via the R statistical computing environment (R Development Core Team, 2007). Documentation for the core packages in the software, including details of the computational algorithms and implementation are given in a special issue of the *Journal of Statistical Software* (`http://www.jstatsoft.org/v24/`). The papers directly related to the computation of the log-likelihood function for ERGM are (Handcock *et al.*, 2008; Butts, 2008; Hunter *et al.*, 2008; Morris, Handcock and Hunter, 2008). Goodreau *et al.* (2008) provides a tutorial documentation of the use of this software for fitting ERGM models.

2. Likelihood-based inference when the network is partially observed. In this section we consider likelihood inference for η in the case where $Y = Y_{obs} + Y_{mis}$ is possibly only partially observed.

We assume that the partial observation mechanism is amenable to the model (1.2). Based on the results in Section 4.1 of Handcock and Gile (2010), the log-likelihood for η is then:

$$\ell[\eta|Y_{obs} = y_{obs}] \propto \kappa(\eta|y_{obs}) - \kappa(\eta)$$

The estimation of the second term in the likelihood is a difficult but well studied problem. The main idea, as reviewed by Hunter *et al.* (2008), is to note that

$$\frac{\kappa(\eta)}{\kappa(\eta_0)} = \mathbb{E}_{\eta_0} \exp\left\{(\eta - \eta_0) \cdot g(Y)\right\},$$

where $\mathbb{E}_{\eta_0}$ denotes the expectation assuming that Y has distribution given by P_{η_0}. Therefore, we may exploit the law of large numbers and approximate the log-ratio by

$$\ell(\eta) - \ell(\eta_0) \approx (\eta - \eta_0) \cdot g(y) - \log\left[\frac{1}{m}\sum_{i=1}^{m} \exp\left\{(\eta - \eta_0) \cdot g(Y_i)\right\}\right], \tag{2.1}$$

where $Y_1, \ldots, Y_m$ is a random sample from the distribution defined by P_{η_0}, simulated using an MCMC routine as described in Section 6 of Hunter *et al.* (2008).

To compute the first term, note that the conditional distribution of Y given Y_{obs}:

$$P_\eta(Y_{mis} = v|Y_{obs} = y_{obs}) \;=\; \exp\left[\eta \cdot g(v + y_{obs}) - \kappa(\eta|y_{obs})\right] \qquad v \in \mathcal{Y}(y_{obs})$$

where $\exp\left[\kappa(\eta|y_{obs})\right] = \sum_{u \in \mathcal{Y}(y_{obs})} \exp\left[\eta \cdot g(u + y_{obs})\right]$. Again note that,

$$\frac{\kappa(\eta|y_{obs})}{\kappa(\eta_0|y_{obs})} = \mathbb{E}_{\eta_0} \exp\left\{(\eta - \eta_0) \cdot g(Y)|\, y_{obs}\right\},$$

where $\mathbb{E}_{\eta_0}(\cdot|y_{obs})$ denotes the expectation assuming that Y has distribution given by P_{η_0} conditional on $Y_{obs} = y_{obs}$. We can approximate the conditional log-ratio by

$$\ell(\eta|y_{obs}) - \ell(\eta_0|y_{obs}) \approx (\eta - \eta_0) \cdot g(y) - \log\left[\frac{1}{m}\sum_{i=1}^{m} \exp\left\{(\eta - \eta_0) \cdot g(Y_i)\right\}\right], \tag{2.2}$$

where $Y_1, \ldots, Y_m$ is a random sample from the distribution defined by $P_{\eta_0}(\cdot|y_{obs})$. These can be simulated using an MCMC routine similar to that for $\kappa(\eta)$ above. Specifically, the conditional distribution of Y given $Y_{obs} = y_{obs}$ is an ERGM on a constrained space of networks, and we can simulate from

it by restricting the proposed networks to the subset of networks that are concordant to the observed data. Specifically, choosing $y_{proposed}$ in equation (12) of Hunter *et al.* (2008) to place positive probability on each network in $\mathcal{Y}(y_{obs})$ and zero mass outside.

Hence the computation of (2) can be based on two separate MCMC samples: the first term by a chain on the complete data and the second by a chain conditional on y_{obs} . So the sampled data situation is typically about twice as difficult as the complete data case. In practice, the estimation can be unstable if the proportion of sampled data is small and the model is near degenerate (Handcock, 2003).

3. Implementation in `statnet`. The computations of the previous section are implemented in `statnet`. They are invoked automatically if the network modeled is partially observed.

To specify that an element of the sociomatrix Y is unobserved, give it a value of `NA`. The `network` package will recognize this as an unobserved value. Otherwise the commands used are identical to that for the fully observed case. Examples are given in Goodreau *et al.* (2008). The package uses the approximation to the log-likelihood to compute approximate standard errors, etc, as for the fully observed case.

4. Two-wave link-tracing samples from a Collaboration Network. In this section we investigate the effect of network sampling on estimation by comparing network samples to the situation where we observe the complete network. Specifically, we consider the collaborative working relations between 36 partners in a New England law firm introduced in Section 1 and analyzed in Section 5 of Handcock and Gile (2010).

The `statnet` code for the original fit is:

```
R> work.fit <- ergm(work ~ edges + gwesp(1)+
                    nodecov("seniority") +
                    nodecov("practice") + match("practice") +
                    match("gender") +
                    match("office"))
```

As discussed in Hunter and Handcock (2006), this model provides an adequate fit to the data, and it is used in the paper to assess the effect of sampling on model fit. We provide the above fit in the `R` workspace `aoas221.fit.RData`. If you use this data in your work, please cite the original source (Lazega, 2001).

We construct all possible datasets produced by a two-wave link-tracing design starting from two randomly chosen nodes (the "seeds"). This adaptive

design is amenable to the model.

For each of these samples we use the methods of Section 2 to estimate the parameters. We can then compare them to the MLE for the complete dataset. For these networks, the MLEs are obtained using `statnet` (Handcock *et al.*, 2003), both for the natural parametrization and for the mean value parameterization (see Handcock, 2003).

The code to do this is in `aoas221.r` followed by `aoas221.summary.r` to produce Figure 2 and Tables 1 and 2. As the mean values are computed by simulation there may be some minor deviations from the values in the tables.

5. Discussion. In this supplement we provide the computational details and code to fit ERGM to networks based on partially observed data (e.g., either from a sampling design or missing due to an amenable model). For additional information and details see `http://statnet.org` (Handcock *et al.*, 2003). The files noted here are available in a single file at `http://statnet.org/aoas221.zip`.

Acknowledgements. The authors would like to thank the members of the UW Network Modeling Group (Martina Morris, P.I.), Stephen Fienberg, and the reviewers for their helpful input.

References.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

Butts, C. T. (2008). **network**: A Package for Managing Relational Data in R. *Journal of Statistical Software* **24**. Available at http://www.jstatsoft.org/v24/i02/

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood calculations (with discussion). *Journal of the Royal Statistical Society, Series B* **54** 657–699.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T. and Morris, M. (2008). A **statnet** Tutorial. *Journal of Statistical Software* **24**. Available at http://www.jstatsoft.org/v24/i09/

Handcock, M. S. (2003). Assessing Degeneracy in Statistical Models of Social Networks Working Paper #39 report, Center for Statistics and the Social Sciences, University of Washington. Available at `http://www.csss.washington.edu/Papers`

Handcock, M. S. and Gile, K. J. (2010). Modeling Networks from Sampled Data. *Annals of Applied Statistics*.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. and Morris, M. (2003). **statnet**: Software Tools for the Statistical Modeling of Network Data Statnet Project `http://statnet.org/`, Seattle, WA R package version 2.0. Available at `http://CRAN.R-project.org/package=statnet`

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. and Morris, M. (2008). **statnet**: Software tools for the representation, visualization, analysis and simulation of social network data. *Journal of Statistical Software* **24**. Available at http://www.jstatsoft.org/v24/i01/

Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics* **15** 565–583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M. (2008). **ergm**: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software* **24**. Available at http://www.jstatsoft.org/v24/i03/

Lazega, E. (2001). *The collegial phenomenon: the social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press, Oxford.

Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley, New York, NY.

Morris, M., Handcock, M. S. and Hunter, D. R. (2008). Specification of Exponential-family Random Graph Models: Terms and Computational Aspects. *Journal of Statistical Software* **24**. Available at http://www.jstatsoft.org/v24/i04/

R Development Core Team, (2007). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0, Version 2.6.1. Available at http://www.R-project.org/

Department of Statistics
University of California - Los Angeles
Los Angeles CA 90095-1554
E-mail: handcock@stat.ucla.edu
krista.gile@nuffield.ox.ac.uk
URL: http://www.stat.ucla.edu/∼handcock
www.nuffield.ox.ac.uk/users/gilek